# Evaluating approaches for the handling of sign reflection in Bayesian latent variable models.

Lihan Chen, Milica Miočević, and Carl F. Falk

McGill University

## Author Note

Corresponding Author: Lihan Chen, lihan.chen@mcgill.ca.

**This is an original manuscript of an article published by Taylor & Francis in Structural Equation Modeling: A Multidisciplinary Journal on 09 Dec 2025, available at: https://doi.org/10.1080/10705511.2025.2588572. The R code and Stan code used in this study are freely available on the Open Science Framework website. No artificial intelligence-generated content tools were used in the production of our study and manuscript.**

**Abstract**

In Markov chain Monte Carlo estimation of Bayesian latent variable models, sign reflection can cause multiple chains to settle onto equivalent but numerically different solutions, resulting in poorly mixed chains and nonconvergence. Sign reflection can be handled using various methods, such as adopting unit loading identification (ULI), assigning range restricted prior distributions, or using a relabeling algorithm. Some statistical software automatically handles sign reflection in the background, e.g., the *blavaan* package in R. We conducted simulations to address the lack of comprehensive studies on such a wide variety of approaches. Our results show that most solutions will work well in confirmatory factor analysis given sufficient sample sizes and good measurement models. However, low scale reliability and poor choice of reference indicator can negatively impact the performance, especially with small sample sizes. In particular, we do not recommend using ULI without additional sign reflection handling for Bayesian latent variable models.

*Keywords:* Bayesian, structural equation model, confirmatory factor analysis, sign indeterminacy

**Evaluating approaches for the handling of sign reflection in Bayesian latent variable models.**

Latent variable models with reflective measurement are often invariant to *sign reflection* in the indicator loadings, which is to say, the model remains equivalent–with the same fit statistics and the same model-implied mean and covariance structures–when some loadings switch signs from positive to negative, or vice versa. This phenomenon, also known as *sign indeterminacy*, is a type of identification problem related to factor reflection in both exploratory and confirmatory factor analysis contexts. Identification itself refers to whether a unique solution exists and can be found for a model under some estimation approach (Hayashi & Marcoulides, 2006). Other types of identification issues involve whether factor loading matrices can be rotated to find an equivalent solution (e.g., including in a confirmatory framework; Millsap, 2001), which also relates to the arbitrary order in which columns appear in such a rotated matrix (for a review of identification issues, see Hayashi & Marcoulides, 2006).

To illustrate sign reflection, suppose latent variable A and B are positively correlated, and a set of indicators positively load onto A, then an equivalent model can be obtained with the same indicators negatively loading onto A, with A now negatively correlated with B. In the empirical interpretation of the model under sign reflection, the label of A is simply reversed, e.g., "high agreeableness is positively correlated with well-being" becomes "low agreeableness is negatively correlated with well-being" under sign reflection. Since the different solutions should lead to equivalent theoretical interpretations, sign reflection is often not a substantive issue for maximum likelihood estimation, where arbitrarily choosing one of the solutions would not affect the conclusion of the analysis. For example, in structural equation model (SEM) software such as *lavaan* (Rosseel, 2012), one loading of each factor may be automatically restricted to positive values in order to present the more easily interpretable solution. Such software programs typically proceed to employ local search algorithms to find the maximum likelihood solution.

In Bayesian Markov chain Monte Carlo (MCMC) estimation, multiple

independent chains are generally used to ensure the posterior space is sufficiently explored. These chains may sample from different local solutions with different factors reflected due to sign indeterminacy, leading to nonconvergence when pooling their draws (Erosheva & Curtis, 2017). The sign reflection problem in Bayesian analysis is related to a broader class of label switching problems that includes cluster labels in Bayesian analysis (Kim, Daniels, Li, Milbury, & Cohen, 2018) and factor labels in Bayesian exploratory factor analysis (Conti, Frühwirth-Schnatter, Heckman, & Piatek, 2014). Methodological and empirical researchers have adopted various approaches to handle sign reflection, such as restricting the signs of the loadings in various ways, including the use of range restricted priors on some or all of the loadings (Chen, Bauer, Belzak, & Brandt, 2022), relabeling the signs in the posterior sample (Erosheva & Curtis, 2017), or switching the signs during Monte Carlo iterations (Merkle & Rosseel, 2018; Merkle, Fitzsimmons, Uanhoro, & Goodrich, 2021).

*Confirmatory factor analysis* (CFA) is used here as an example in order to discuss sign reflection in more detail, with generality retained under most latent variable models with "reflective" measurement models. In CFA, indicators are a function of their underlying factor(s). For example, a participant's score on an indicator may be $x = a + \lambda F + e$, where $F$ is the factor score for that participant for the factor underlying that indicator, $\lambda$ is the factor loading, $e$ is the random measurement error, and $a$ is the indicator intercept. Since the factor score $F$ is not observed except indirectly through the indicator $x$, the scale for $F$ must be fixed in some way for the CFA model to be identifiable (see Millsap, 2001). In practice, a latent variable model is typically identified through one of two approaches. The *unit variance identification* (UVI) approach fixes the variance of all factors to 1. The *unit loading identification* (ULI), fixes one indicator loading in each factor to 1, thus scaling the factor score to a *reference indicator* (Bollen, Lilly, & Luo, 2022). Sign indeterminacy is primarily an issue under UVI, where the indicator score $x$ would remain the same, if for example, the values in $\lambda$ and $F$ were both reflected around 0 – suppose $\hat{\lambda} = -\hat{\lambda}'$ and $\hat{F} = -\hat{F}'$, then $\hat{\lambda}'\hat{F}'$ and $\hat{\lambda}\hat{F}$ will yield the exact same value. This indeterminacy in turn can cause multiple chains to

sample from reflected equivalent solutions, resulting in poorly mixed chains with poor convergence and invalid summary statistics for the Bayesian estimation of the loading $\lambda$.

## Strategies for Sign Reflection

Quantitative and empirical researchers conducting Bayesian analysis on latent variable models adopt a wide range of approaches for the handling of sign reflection. Most of these approaches can be broadly divided into the following categories: 1) unit loading identification (ULI); 2) range restricted priors, where the priors of some or all loadings are restricted to the positive range to avoid sign reflection; 3) posterior relabeling, such as the algorithm by Erosheva and Curtis (2017), which takes the posterior draws from multiple chains to switch the signs of reflected solutions; 4) in-iteration relabeling, used by *blavaan* (Merkle & Rosseel, 2018) to switch the signs of loadings during the MCMC sampling to ensure the solutions have the expected signs. In what follows, we provide an overview of these approaches.

### *Unit Loading Identification*

With ULI, because one reference indicator loading of each factor is chosen to be fixed to 1, the sign of the solution to the factor scores is constrained to be in agreement with the sign of that indicator. In theory, this should make a latent variable model identified using this strategy resilient to issues caused by sign reflection. However, the ULI has several downsides. First, if the chosen reference indicators are poor measures of the factors, inefficiency may permeate other aspects of the model, such as instability in the scaling of the factor scores and unreasonably large and highly varied loadings on other indicators. Second, if an indicator has low reliability, a large portion of the sampling distribution of its loading (had it been estimated) may be on the other side of 0. Thus, fixing reference loadings to 1 may cause the factor scaling itself to change signs from sample to sample. Lastly, in many applications of the latent variable model, it is important to establish *measurement invariance*, i.e., the scale works the same way across different groups. ULI may not be desired in such a scenario, because the identification strategy already assumes that the reference indicators are invariant across groups. For these reasons, it is important to explore solutions to sign reflection in

Bayesian latent variable models not identified using ULI.

### *Range Restricted Priors*

In Bayesian estimation, loadings are often given diffuse normal priors centered around 0. For example, the default prior for loadings in *blavaan* is $\mathcal{N}(0, 10^2)$ (see Merkle & Rosseel, 2018; Smid & Winter, 2020). The range restricted priors approach handles sign reflection in UVI latent variable models by ensuring that one or several loadings for each factor are positive by assigning them priors that only have density in the positive range. One common approach is through the use of *truncated normal priors*, specifically half-normal priors, which are normal priors restricted to only positive values. For example, in *item response theory* (IRT), a categorical variable counterpart to CFA, half-normal priors $\mathcal{N}^+(0, 5^2)$ are sometimes given to the slopes (similar to loadings in CFA) (Chen et al., 2022). Alternatively, for researchers who prefer to avoid potential improper priors with truncated priors, a more natural choice could be lognormal priors, which are naturally bounded at 0 (Martin-Fernandez & Revuelta, 2017).

Ulitzsch, Lüdtke, and Robitzsch (2023) studied small sample Bayesian estimation of SEMs. In their study, Bayesian models were used to estimate the standardized loadings and latent variable correlations through the well-established *model-implied covariance matrix* (Jöreskog, 1970) for SEM, using the sample covariance matrix of indicators rather than individual participant data as input. By parameterizing the model with standardized loadings rather than unstandardized loadings and assuming very simple structure (each indicator loads on only one factor at a time), each loading must fall into the range of $[-1, 1]$. Thus, sign reflection can be handled under this approach by assigning uniform priors $Uniform(.01, .99)$ to the first loading of every factor to ensure it is positive, and the remaining loadings can be assigned uniform priors $Uniform(-.99, .99)$.

Despite the intuitive appeal of range restricted priors, researchers are sometimes advised against using this approach. When loadings are small or the sampling distribution has high variability, the posterior distributions for loadings are likely to have some probability density in both positive and negative regions if unrestricted

priors were used. Since priors bounded above zero cannot produce these posteriors samples, they are unable to properly estimate these loadings. Instead, approaches such as relabeling, which are in theory more robust to problems such as low reliability and poor reference indicators, are often recommended as alternatives (Merkle, Ariyo, Winter, & Garnier-Villarreal, 2023).

### *Posterior Relabeling*

When sign reflection occurs without additional estimation issues, the reflected posteriors of each parameter (e.g. loadings, factor covariances, path coefficients) are theoretically sampled from distributions that are identical but for their reflected signs across multiple chains. This allows us to pursue a post hoc relabeling approach, first performing the multiple chain Bayesian estimation with UVI—temporarily tolerating sign reflection—then multiplying the appropriate loadings and factor covariances (or path coefficients) by -1. Erosheva and Curtis (2017) proposed a *relabeling* algorithm for Bayesian factor analysis. The algorithm leverages the assumption that the posterior samples for a loading from multiple chains are pooled together, they should form a roughly normal distribution when no sign reflection occurs. This assumption becomes more reasonable as sample sizes increases as posterior distributions will approach normality. Thus, this approach examines the posterior samples across all chains to choose a set of sign changes that results in a pooled posterior distribution for each parameter, which best fits a normal distribution. The sign changes are chosen by first assigning a set of random initial sign changes, then iteratively finding new sign changes that maximize a normal density for the posterior of each loading. Compared to range restricted priors, relabeling has the advantage of being able to appropriately obtain posteriors for loadings that overlap with zero. However, the success of this approach is somewhat contingent on the pooled posterior across reflected chains for each parameter having reasonably well separated modes (Erosheva & Curtis, 2017). If the modes are not well separated, for example, if the loadings are too close to zero, or if other estimation issues occur such as the chain being stuck in local extrema, the posterior relabeling approach may not be able to appropriately handle sign reflection.

### *In-Iterations Relabeling*

The R package *blavaan* (Merkle & Rosseel, 2018) handles sign reflection using an *in-iterations relabeling* approach (Merkle et al., 2021). Similar to some range restricted priors approaches, in-iterations relabeling restricts one "reference" loading in each factor to be positive, allowing remaining loadings to be freely estimated. Rather than placing range restricted priors on the reference loadings, this approach ensures they are positive by checking their sampled values during the MCMC iterations. At any iteration, if a draw for a reference loading is negative, that reference loading is multiplied by -1 before the MCMC estimation continues, as are other loadings for that factor and any associated covariances and path coefficients. Subsequent iterations thus typically get back on track to a solution in which the sign of the referent loading is positive and ensuring sign agreement across chains. While this is the automatic sign reflection handling in *blavaan*, it is not documented in detail and its impact has not been thoroughly studied[1].

## Current Study

To our knowledge, there has not been a comprehensive study on sign reflection handling in Bayesian latent variable models that summarizes and evaluates the effectiveness of such a wide variety of approaches. Here we conduct a simulation study to evaluate the performance of a range of extant approaches, from ULI, Erosheva and Curtis (2017)'s relabeling algorithm, *blavaan*'s in-iterations relabeling method, to several range restricted prior approaches, including Ulitzsch et al. (2023)'s marginal likelihood approach placing restricted priors on standardized loadings and factor corelations. We used a three factor CFA for the simulation, as CFA with two to three factors are very common psychology and other social sciences. Additionally, a three factor CFA is equivalent to a latent variable mediation model, which is also a commonly used model in applied research and a focus of several recent simulation studies evaluating Bayesian methods (e.g., Ulitzsch et al., 2023, Liu, Heo, Ivanov, & Depaoli,

[1] We would like to thank Dr. Ed Merkle for making *blavaan* open source, as well as clarifying its sign handling feature in private correspondence

2025, Sun, Zhou, & Song, 2021, (Liu, Heo, Depaoli, & Ivanov, 2025)). For the simulation, we varied the overall reliability of the scales (high vs. low), the strength of the reference loadings (good vs. poor), and the sample size (high vs. low), to provide an evaluation under a range of plausible scenarios.

## Methods

The simulation study was conducted using the R programming language (R Core Team, 2019) and the Stan software (Stan Development Team, 2025). R code and Stan code were adapted from Ulitzsch et al. (2023) and Erosheva and Curtis (2017), respectively, for the marginal likelihood approach with restricted priors on standardized loadings and the relabeling strategies. The *blavaan* package was used for the in-iterations relabeling strategy (Merkle & Rosseel, 2018). All the code used to simulate the data and perform the analyses are available on OSF[2].

### Data Generating Model

The data generating model was a CFA with three factors and four indicators per factor, $\mathbf{X}_{ij} = a_{ij} + \lambda_{ij}\mathbf{F}_j + \mathbf{e}_{ij}$, where $j \in [1, 2, 3]$ for the three factors, $i \in [1, 2, 3, 4]$ for the four indicators. $\mathbf{X}_{ij}$, $\mathbf{F}_j$, and $\mathbf{e}_{ij}$ were random normal variables each representing indicator scores, factor scores, and measurement errors, respectively. $\lambda_{ij}$ represent indicator loadings, while $a_{ij}$ are the indicator intercepts and were set to 0. Correlations between the factors were set to be .3. Measurement error was assigned to each indicator so that they were standardized in the population, i.e. the variance of the measurement error was set to $var(\mathbf{e}_{ij}) = 1 - \lambda_{ij}^2$.

For our study design, we wanted to evaluate as many approaches as possible, although many of them can be somewhat computationally intensive. Therefore we limited the number of data conditions, with three fully crossed factors of sample size, reliability, and reference loading quality, leading to 2 (small vs. medium sample size) × 2 (low vs. high reliability) × 2 (good vs. poor reference loadings) = 8 conditions in total for each of the 7 approaches.

For the sample size conditions, we chose $N = 100$ and 200. The sample size

[2] Available at `https://osf.io/392ng/?view_only=6e35f8676d87439ca67286ad9ca76dc0`

$N = 100$ was chosen to examine how methods can breakdown under a small, but nevertheless realistic, sample size for CFA in psychological research. The $N = 200$ is typically deemed as a minimum sufficient sample size under the same context. In the factor analysis literature, 0.3 is often regarded as the minimal loading for the inclusion of an item in a scale, whereas 0.7 is often regarded as an "acceptably good" loading. For the low reliability conditions, loadings were set to 0.7 for half the items and 0.3 for the other half of the items in each factor, for a congeneric reliability of $\omega = .585$ per factor. This allowed us to set the reference indicator loading to 0.7 for the "good" reference indicator condition, and 0.3 for the "poor" reference indicator conditions. For the high reliability conditions, we following the same criteria, choosing loading values of 0.6 and 0.8, which fall on the two sides of the "acceptably good" 0.7, as the loadings for the items, with a congeneric reliability of $\omega = .797$. Thus, the reference indicator loading was set to 0.8 for the "good" reference loading condition, and 0.6 for the "poor" reference loading condition, for the high reliability conditions.

We simulated samples of the 12 indicator values (4 per factor for 3 factors) under a multivariate normal distribution, where the indicator covariance is the model-implied covariance matrix of CFA (e.g. Jöreskog, 1970; see code on OSF for technical details) derived from the model defined above, and the means are all set to 0. In this case, this would be equivalent to assuming both factor scores are multivariate normal and measurement errors are normal. For each condition, 500 data sets were independently generated for the following analyses.

**Analysis**

For the analysis, we included 7 approaches in total. 1) We included the *unit variance identification* (UVI) as a control condition to establish the performance of CFA in the absence of sign reflection handling. Due to sign reflection, we expected this approach to perform poorly, but it was unclear to us precisely how poorly it would perform, and how much of an improvement the other methods would be. 2) For comparison, we also included *unit loading identification* (ULI). We expected ULI to outperform UVI, but it could still yield poor performance when the reference indicator

was poor. 3) We also included an approach directly employing the *blavaan* package, which employs in-iterations relabeling automatically to handle sign switching. For convenience, we refer to this approach simply as "blav". 4) Furthermore, we included Erosheva and Curtis (2017)'s posterior relabeling algorithm which handles sign reflection in a more principled manner; we refer to this approach as "relab". Finally, for range restricted priors, because there is such a wide range of possible implementations, we limited the study to include three alternatives. 5) Firstly, we included Ulitzsch et al. (2023)'s approach that placed uniform priors on standardized parameters using a marginal likelihood parameterization, which we will refer to as "margstd". 6) Secondly, we included a straightforward, more "brute force" approach of simply assigning truncated halfnormal priors to all parameter loadings, which we refer to as "alltrunc". 7) Lastly, we wanted to include an approach that would be more minimal, placing fewer restrictions and using naturally bounded priors rather than truncated priors. To this end, we included a condition where we placed a lognormal prior on the reference indicator of each factor, and we refer to this condition as "logref".

For 1) UVI, 2) ULI, 4) relab, 6) alltrunc, and 7) logref approaches, we constructed Bayesian analysis models in Stan using conditional likelihoods, implemented in *R* v4.4.0 (R Core Team, 2019) and Stan v2.32.2 (Stan Development Team, 2025), with the *R* package *cmdstanr* v0.8.1 and CmdStan v2.36.0. The 4) magstd approach was also constructed in Stan, except with code adapted from Ulitzsch et al. (2023). For the 3) in-iterations relabeling approach (blav), we directly used the *blavaan* package to construct a CFA model under unit variance identification following *lavaan* syntax. The *blavaan* package then automatically converts *lavaan* CFA syntax into a marginal likelihood Bayesian model (Merkle et al., 2021). For all 7 approaches, 5,000 burn-in iterations and 10,000 sampling iterations were performed with 3 chains for each of the 500 replications.

### *Unit Variance Identification*

For 1) UVI, we defined the likelihood of each indicator as $\mathcal{N}(a_{ij} + \lambda_{ij}\mathbf{F}_j, \epsilon_{ij})$, where $\epsilon_{ij}$ was the standard deviation of the measurement error $\mathbf{e}_{ij}$. To facilitate more

direct comparisons, we assigned priors to the model parameters similar to the defaults in *blavaan*. The three sets of factor scores $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ were assigned a multivariate normal prior with a covariance of $\mathcal{LKJ}(1)$ (Lewandowski, Kurowicka, & Joe, 2009), which always yields 1s as the diagonal elements on the covariance matrix, providing unit variance identification for our model. Each $\lambda_{ij}$ was assigned a prior of $\mathcal{N}(0, 10^2)$, each $a_{ij}$ was given a prior of $\mathcal{N}(0, 32^2)$. Each measurement error standard deviations $\epsilon_{ij}$ was assigned a prior of $Gamma(1, .5)$. To help keep the comparisons fair, starting values were chosen to be similar to those in *blavaan* when performing CFA on variables of a similar scale. Specifically, the loadings were assigned random starting values from $Uniform(.8, 1.2)$, measurement error standard deviations from $Uniform(.4, .6)$, item intercepts from $Uniform(-0.05, 0.05)$. The factor correlation matrix was given the identity matrix as the starting value.

***Unit Loading Identification***

The 2) ULI model was identical to the UVI model except for the following aspects. First, the first loading of each factor was fixed to 1, rather than freely estimated. Second, the covariance between factor scores was specified as $\tau\Omega\tau$, where $\Omega$ was the correlation matrix given a prior of $\mathcal{LKJ}(1)$, while $\tau$ was a diagonal matrix with the standard deviations of $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ on its diagonal, which are each in turn given a prior of $Gamma(1, .05)$. This frees the scale of the indicators so their variances are no longer fixed to 1 and are freely estimated. Indicator loadings are no longer on a standardized scale under ULI, so we set the initial value of the loadings to a wider range of $Uniform(.5, 1.5)$. Since the data generating model adopts a scale based on standardized factors and indicators (i.e., indicators and factors have standard deviations of 1), the unstandardized loadings from the ULI approach are not on the appropriate scale for the evaluation of the loading estimates. Therefore, at each iteration of the Bayesian Monte Carlo estimation, we computed partially standardized loadings (with respect to the variance of the factors only) from the unstandardized loadings and factor standard deviation. The posteriors of these partially standardized loadings are used for the Bayesian estimation.

### *In-Iterations Relabeling*

For the 3) in-iterations relabeling strategy, we specified a three factor CFA with standardized factors, free loadings, and free factor correlations for the `bcfa` function in the *blavaan* package with default priors (code available on OSF). The default priors were $\mathcal{N}(0, 10^2)$ for the loadings, $\mathcal{LKJ}(1)$ on the correlations, $Gamma(1, .5)$ for the standard deviation of measurement errors, and $\mathcal{N}(0, 32^2)$ for the indicator intercepts. Internal functions of *blavaan* first convert the model specified in the *lavaan* syntax into Stan code, before *blavaan* automatically uses Stan to perform the Bayesian estimation (Merkle et al., 2021). It is worth noting that *blavaan* also uses a marginal likelihood approach, similar to Ulitzsch et al. (2023), but the sign reflection handling is entirely differently. Inside the Stan code for *blavaan*, a generated quantity block is used to, at each iteration, multiply any sampled value of reference loadings below 1 by -1, which may entail also reversing the sign of other associated loadings, path coefficients, or factor covariances. This forces the reference indicator loadings to always be positive. The default number of burn-in and sampling iterations are set to 500 and 1,000 in *blavaan*, but we increased them to 5,000 and 10,000 to match our other analyses.

### *Posterior Relabeling*

The 4) posterior relabeling approach was performed modifying a function provided by Erosheva and Curtis (2017). First, Bayesian CFA was conducted in Stan using the UVI approach with three chains as usual. Next, let $\mathbf{\Lambda}$ be a $p \times q$ matrix of factor loadings, and $\boldsymbol{\nu}^{(t,c)}$ be a vector of length $q$ that contains sign changes associated with the $q$ factors at iteration $t = 1, \ldots, N$ for chain $c = 1, \ldots, C$ of the MCMC samples. The goal is to make changes to to them to minimize a function equivalent to the following loss:

$$\sum_{t=1}^{N}\sum_{c=1}^{C}\min_{\boldsymbol{\nu}^{(t,c)}}\left\{-\sum_{i=1}^{p}\sum_{j=1}^{q}\mathbb{I}\left(\lambda_{ij}^{(t,c)}\neq 0\right)\log\left[f\left(v_j^{(t,c)}\lambda_{ij}^{(t,c)}; m_{ij}, s_{ij}^2\right)\right]\right\}$$

where $m_{ij}$ and $s_{ij}^2$ are the interim posterior means and variances, respectively, for $\lambda_{ij}$, $f(\cdot; m, s^2)$ is a normal density, and $\mathbb{I}\left(\lambda_{ij}^{(tc)} \neq 0\right)$ is an indicator function that is 1 only for loadings that were not fixed to zero, and zero otherwise. After randomly initializing

the sign change vectors, $\boldsymbol{\nu}^{(t,c)}$, the algorithm attempts a new sign switch for one of the factors to see if it improves the loss function. If a sign switch improves loss, it is accepted as the new sign switch, and the process is repeated until no new proposed sign switches further improve the loss.

Since the original relabeling function by Erosheva and Curtis (2017) did not include the relabeling of factor covariances, we implemented a simple extension which takes the sign switches for loadings from the relabeling function and computes the corresponding sign switches on factor correlations. Specifically, for each factor correlation, if one and only one of the factors switches sign due to the relabeling algorithm, the sign of the corresponding factor correlation also switches (code available on OSF). After the sign switches for the loadings and factor correlations are obtained, the posteriors for each of these parameters are multiplied by the sign switches to complete the relabeling, and summaries of these relabeled posteriors are used in the estimation of the parameters.

### *Range Restricted Priors*

For the range restricted priors approaches, 6) all trunc and 7) logref used analysis models similar to the UVI model with slight modifications. With the alltrunc approach, all loadings were instead assigned $\mathcal{N}(0, 5^2)$ and restricted to be greater than 0 in Stan, effectively assigning halfnormal priors $\mathcal{N}^+(0, 5^2)$ to all loadings. With the logref approach, the reference/first loading of each factor was instead assigned a prior of $Lognormal(0, 1^2)$ to ensure that it would be positive.

The 5) margstd approach from Ulitzsch et al. (2023)[3] did not parameterize the indicator scores or the factor scores. Instead, the means and covariances among the indicators were modeled directly. First, let the model-implied covariance matrix be $\Sigma$. Although, $\Sigma$ is unstandardized, it is computed from standardized loadings and factor correlations, rescaled by the standard deviations of the indicators. The factor

[3] The original code specified structural regressions with 3 indicators per latent variable. We simplified it to a CFA model with 4 indicators per factor and applied some bug fixes in the original code; code and comments available on OSF.

correlations were assigned $Uniform(-.99, .99)$ priors. The first standardized loading of each factor was assigned a $Uniform(.01, .99)$ prior, restricting it to be positive. The remaining standardized loadings were assigned $Uniform(-.99, .99)$ priors. The indicator standard deviations were assigned $\mathcal{N}(0, 10^2)$ priors. The scatter matrix of the data $S$, i.e., the covariance matrix of the indicators multiplied by $N - 1$, where $N$ was the sample size, was assumed to follow a $Wishart(N - 1, \Sigma)$ distribution. The means of the indicators were assumed to follow a $\mathcal{MVN}(\mathbf{a}, \frac{\Sigma}{N})$ distribution, where $\mathbf{a}$ were the indicator intercepts assigned $\mathcal{N}(0, 10^2)$ priors. Following Ulitzsch et al. (2023), we sampled the initial values of all loadings from $Uniform(.01, .99)$, and relied on Stan defaults for start values for other parameters. To compare the standardized loading estimates produced by margstd to the unstandardized loading estimates produced by other approaches in the study, we multiplied each loading with the estimated standard deviation of its corresponding indicator.

## Evaluation

For each method in each condition, the summary of the posteriors was used to evaluate the performance of the approach. The primary evaluation method was convergence. When sign reflection occurs in Bayesian latent variable models, the posteriors across multiple chains would typically fail the convergence criteria. With the converged replications, the means of the posteriors were used as the parameter estimates, and their mean and standard deviations across the replications were used to compute the average bias and the root mean square error (RMSE). Finally, we checked the 95% credible interval coverage of the posteriors.

### *Convergence*

Four converge criteria were examined, including the *potential scale reduction factor* ($\hat{r}$), Monte Carlo standard error (MCSE), and bulk and tail *effective sample size* (ESS) (Stan Development Team, 2025). $\hat{r}$ is the ratio of the between chain variance and the within chain variance (Gelman & Rubin, 1992). We only treated a run as converged under this criterion if $\hat{r} < 1.1$ for every loading, factor correlation, and measurement error. The MCSE is the standard error of the posteriors, and it should be small relative

to the standard deviation of the posteriors (MCSD). We used the threshold MCSE/MCSD $< .1$ for convergence. ESS-bulk is an estimate of the average number of independent draws in the posterior sample (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021). ESS-tail is an estimate of independent draws for the 5% tails of the posteriors (Vehtari et al., 2021). ESS bulk and tail should be at least 100 per chain as an indication of convergence (Stan Development Team, 2025; Vehtari et al., 2021).

### *Bias, RMSE, and Coverage*

The remaining statistics used for evaluation were computed from the converged replications for each condition and each approach. To simplify the notation across all parameters (loadings, factor covariance, and so on), we use $\theta$ to denote some parameter in the model, and $\hat{\theta}_i$ to denote the sample estimate of that parameter in the $i$th replication of the simulation study. $\hat{\theta}_i$ was computed from the mean of the posterior distribution. The average relative bias is given by $\frac{1}{R}\Sigma_{i=1}^{R}\frac{\hat{\theta}_i-\theta}{\theta}$, where $R$ is the number of converged replications. RMSE is given by $\sqrt{\frac{1}{R}\Sigma_{i=1}^{R}(\hat{\theta}_i-\theta)^2}$. In the case an estimate is unbiased, lower RMSE indicates better efficiency. To make the results more interpretable, we computed the relative RMSE for each approach compared to the in-iterations relabeling approach used by *blavaan* (blav). For example, the relative RMSE of the alltrunc approach is $1-\frac{RMSE_{alltrunc}}{RMSE_{blav}}$. A value of 0.10 indicates the RMSE is 10% larger compared to blavaan, while a value of -0.10 indicates the RMSE is 10% lower compared to blavaan. The 95% coverage, with the exception of the univariate approach, is given by the percentage of converged replications where the 2.5% to 97.5% equal-tailed interval of the posterior distribution contains the true population value of the parameter $\theta$.

## Results

### *Convergence*

The proportion of converged runs out of 500 replications based on each criterion can be seen in Figure 1. Consistent with expectations due to sign reflection, UVI practically never converged under any criterion, with convergence below 3% according to any criterion. Although ULI performed poorly based on the ESS criteria especially

when reliability was low, it could achieve reasonable convergence with $\hat{r}$ and $MCSE$ criteria under high reliability. Somewhat surprisingly, the more "natural" approach with lognormal priors on reference loadings (logref) did not meaningfully improve convergence, yielding between 0.05% and 30% convergence on the $\hat{r}$ and MCSE ratio criteria, and below 16% on ESS criteria. In contrast, assigning truncated priors to all loadings led to generally acceptable convergence with $\hat{r}$ and $MCSE$, but yielded lower ESS under low reliability conditions, around 55% to 62%. The relabeling approach yielded slightly better performance than the alltrunc approach, but the performances for these two methods were similar. The two marginal likelihood approaches, one using range restricted uniform priors on the reference indicators (margstd), and one using the in-iterations relabeling strategy (blav), both performed best overall, although margstd had poorer convergence when the sample size was small, reliability was low, and the reference indicator was poor.

### *Bias, RMSE, and Coverage*

To evaluate the performance of the parameter estimates, we examined the aggregated relative bias, relative RMSE, and coverage. The number of converged runs were too low for UVI and logref for meaningful summaries, so they were excluded from the performance evaluation. Further, to include more available replications, especially for ULI, ESS bulk and tail were not used as convergence criteria for the summaries below. This means that for these analyses, replications were included as long as both $\hat{r} < 1.1$ and $MCSE/MCSD < .1$. The number of converged runs under this set of criteria is shown in Figure 2. Alternative analyses using all four criteria for convergence are available on OSF, which yielded largely the same pattern of results, except under low reliabilty conditions where ULI did not always provide meaningful performance summaries due to low number of converged replications.

Figure 3 shows the average relative bias over the converged replications. While ULI yielded overall acceptable performance, it underestimated the reference loading in most conditions. First, when the sample size was small (100), ULI only yielded unbiased estimates (below 10% absolute relative bias) when both the reliability was

high and the reference indicator was good. Second, even when the sample size was moderate ($N = 200$), under the low reliability and poor reference indicator condition, ULI yielded a -0.26 relative bias. Lastly, under the condition of small sample size, low reliability, and poor reference indicator, the relative bias was very large at -0.40. The in-iterations relabeling marginal approach with *blavaan* (blav) yielded the best convergence overall, but it also resulted in slightly larger negative relative bias for the factor covariances compared to the other approaches, primarily when the sample size was small and reliability was low. Other approaches also somewhat underestimated the factor covariances, especially when the reliability is low, but these were not particularly large absolute relative biases (close to 10% or below). The *blavaan* approach was also more likely to underestimate the loadings when the sample size was small, the reference indicator was poor, and reliability was low, but the absolute relative bias was also reasonably small (between 5%-10%). Except for the small bias in *blavaan* and the notable bias in the reference indicators for ULI, loading estimates were largely unbiased under all other approaches.

In Figure 4, we present the relative RMSE of each condition in comparison to *blavaan*. Generally, the relative RMSE reflected the bias in Figure 3, showing practically little difference in the efficiency between any two methods when they both yielded unbiased estimates. For example, ULI yielded larger RMSE in reference loadings except when the reliability was high and the reference indicator was good, reflecting its biased estimates in other conditions. Similarly, most conditions yielded RMSE lower compared to *blavaan* under the condition where sample size was small, the reliability was low, and the reference indicator was poor. The only exception is that the margstd approach tended to yield slightly lower RMSE than all other approaches, regardless of conditions.

Finally, all approaches showed good coverage generally (Figure 5). The only poor coverage occurred under ULI under low reliability, with either low sample size or poor reference indicators, where the approach also struggled with convergence and unbiased estimation.

## Discussion

Our study examined a range of extant approaches in the literature for handling sign reflection in Bayesian latent variable models, in particular with a three factor CFA model. In general, most approaches performed very well, even when the reliability of the scales were fairly low ($\omega = .585$), the reference indicator was quite poor ($\lambda = .3$), and the sample size was quite small ($N = 100$). While the UVI approach consistently encountered sign reflections, it was encouraging to find that most extant approaches could handle the problem reasonably well.

A somewhat surprising finding was that assigning lognormal priors to the reference indicator (logref) did not appropriately handle sign reflection under the scenarios studied. With logref, we were trying to include a condition in which we impose fewer constraints on the parameters and choose more "natural" priors by avoiding improper priors. The lognormal priors could be suitable to other parameters, and perhaps with larger sample sizes, but when assigned to just a single loading per factor in a CFA model under our study conditions, convergence was hardly ever achieved.

With the alltrunc approach, we assigned truncated, halfnormal priors to all loadings to ensure that loadings are positive. This led to good convergence overall, except for low ESS bulk and tail when reliability was low, where the convergence rates based on each of those criteria were around 60%. When reliabilities, and hence the loadings, were low, the sampling of the posteriors was more likely to abut the truncated boundary of 0, which may have increased the autocorrelation in the posteriors. However, this did not seem to have an appreciable impact on $\hat{r}$ and the MCSE ratio, or notably affect the parameter estimates in terms of bias, RMSE, or the coverage of the 95% credible interval. The alltrunc approach presents a curious contrast with the logref approach, where a more forceful approach outperformed one that could be deemed less restrictive.

The margstd approach based on (Ulitzsch et al., 2023), similar to the alltrunc approach, also imposes improper priors with restricted range on the loadings, but it is different in three primary ways: 1) The Bayesian model is parameterized using a

marginal likelihood approach, modeling the mean and covariance structure without needing the factor scores; 2) The priors are placed on standardized loadings with uniform distributions; and 3) Only one loading on each factor is constrained to be positive. This resulted in good convergence in general, as well as unbiased parameter estimation with good efficiency, along with coverage. Lastly, the marginal likelihood approach also has the advantage of having a faster computation time, as fewer parameters are involved in the model.

The posterior relabeling approach (relab) by (Erosheva & Curtis, 2017) represents one of the more theoretically sound approaches, where signs are switched in the posteriors so that each posterior distribution appears normal. Somewhat surprisingly, it showed practically identical performance to alltrunc, having somewhat low ESS bulk/tail when scale reliability was low, but performed very well otherwise. It may be the case that when reliability is low, the MCMC samples with no constraints do not necessarily settle into one of two normal distributions with either a positive or negative mean that are reflections of each other around zero. Instead, the samples may waver between different solutions, resulting in a posterior that cannot simply be fixed by finding the best sign switch.

The *blavaan* also implements a marginal likelihood approach (blav) similar to (Ulitzsch et al., 2023), but it uses an in-iterations relabeling approach with priors without range restriction. This resulted in the best convergence overall, albeit with a slightly higher negative bias on the factor covariance. The run time in *blavaan* tends to be longer than in all other approaches, often twice as long in our simulation, possibly due to the overhead of the more automated software.

Overall, it is a reassuring finding that despite the differences in a wide range of approaches available for the handling of sign reflection, most of them yielded decent performance. Out of all the approaches, ULI is the least recommended for the handling of sign reflection because it performed very poorly under low reliability and poor indicators and only yielded comparable performance to other approaches under ideal conditions. For researchers who wish to use a more accessible and user friendly

approach, *blavaan*'s automatic sign reflection handling will generally be able to avoid nonconvergence and provide good estimates, but there may be a slight trade off in increased bias and lower efficiency when the sample size is small and the scales are poor. For researchers more willing to perform some custom coding in Stan, R, and other statistical coding languages, we found that posterior relabeling, uniform priors on standardized parameters, and truncated normal priors are all able to yield good performances, with a computational speed advantage and potential slight advantage in efficiency for the standardized parameterization.

**Limitations and Future Directions**

To our knowledge, our study is the first to directly compare the performance of a wide range of sign switching handling approaches, and as such, we only focused on cases without assumption violations and model misspecifications. Additionally, we used a common 3-factor CFA model, and limited ourselves to relatively realistic scenarios when selecting simulation parameters for low reliability and poor reference indicator conditions. While we expect many results to generalize to many of these cases, for example, CFA with a mix of negative and positive loadings for the same factor, researchers need to be cautious about applying the incorrect range-restricted prior (e.g., alltrunc) for some loadings. Thus our findings may not represent what could happen under more egregious conditions. For example, loadings of 0.1 may cause more difficulty for handling sign reflection. Likewise, there is sometimes an interplay between sample size, the number of factors and factor correlations, and the number of indicators. We conjecture that more nontrivially correlated factors and/or decent indicators may sometimes help stabilize estimation, provided that sample size is not too small. Conversely then, too few indicators per factor may yield more difficulties. Future studies may also explore more specific ways the sign switching handling methods could be affected by assumption violations, such as different nonnormal distributions in the factor scores and/or measurement errors, a mix of positive and negative loadings, misspecified models with omitted cross loadings or correlated errors, etc.

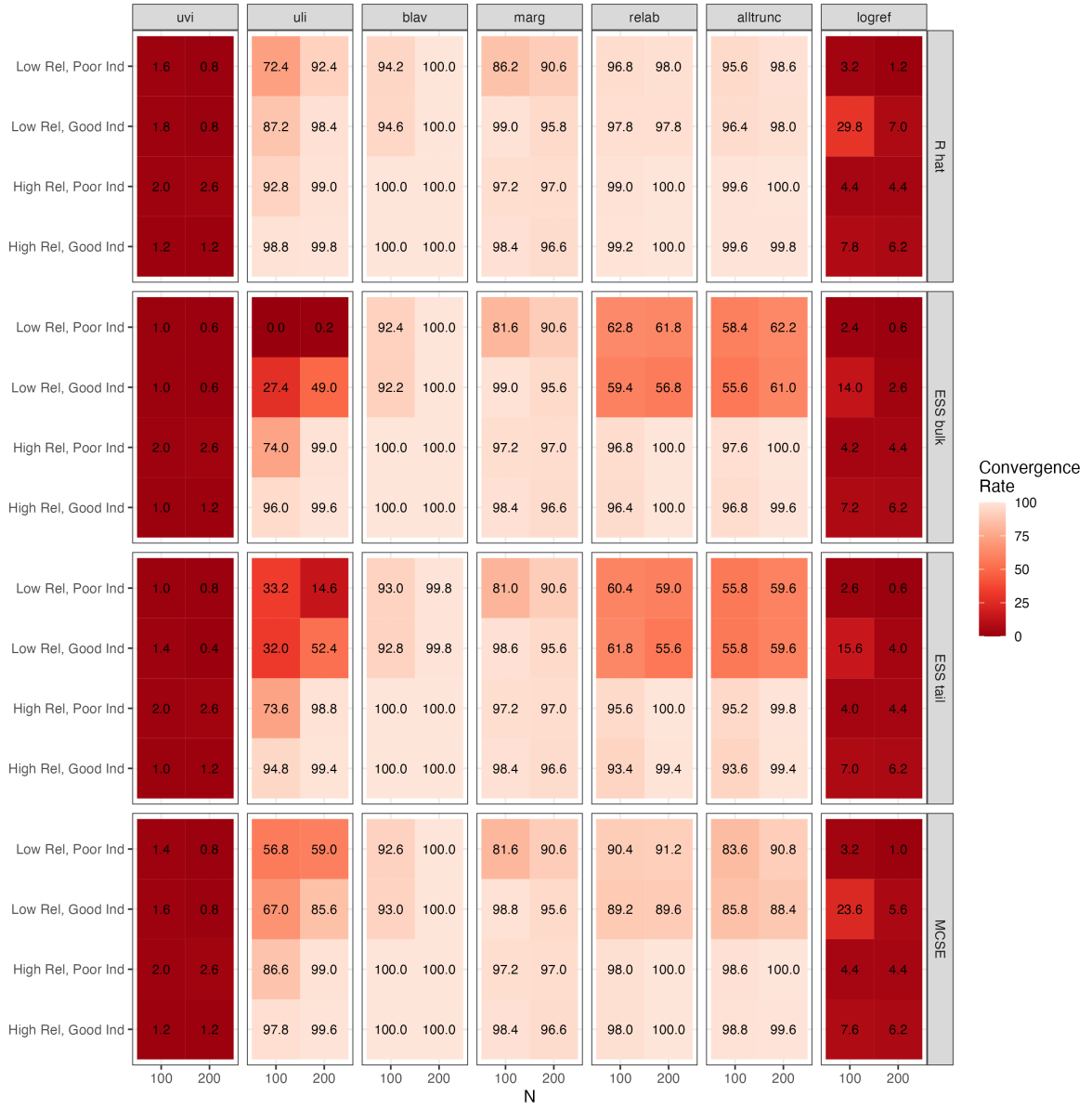
An alternative software program for Bayesian estimation that we did not study

is M*plus* (Muthén & Muthén, 2017). M*plus* does not explicitly document sign switching handling in its User's Guide and does not provide open source code, but according to a webnote the software attempts to avoid between-chain label switching by running 50 identical iterations for each chain (Asparouhov & Muthén, 2010) and, according to later presentation slides on the new features in M*plus* 7 (Muthén & Asparouhov, 2012), the software automatically handles sign switching by constraining the sum of all loadings in each factor to 1. This constraint, unlike alltrunc, allows for negative loadings, but ensures that the positive loadings are greater in magnitude in each factor, to prevent sign switching within chains and encourage sign agreement across multiple chains. While this constraint is weaker than alltrunc, given that M*plus* assigns starting values of 1 to loadings of continuous indicators by default, it should offer comparable performance. As with all software, researchers are advised to examine trace plots to ensure the chosen solution for label switching successfully prevented the issue.

Lastly, many approaches we studied also come with a wide range of possible implementations based on researchers' decisions. For example, our implementation of assigning lognormal priors to reference loadings did not pan out for our model, but a modified version of an approach following the same theoretical idea may work. A full exploration of potential implementations is outside of the scope of the study, which aims to provide an overview of a broad range of approaches, and we leave more specialized investigations to future studies.

## Figure 1

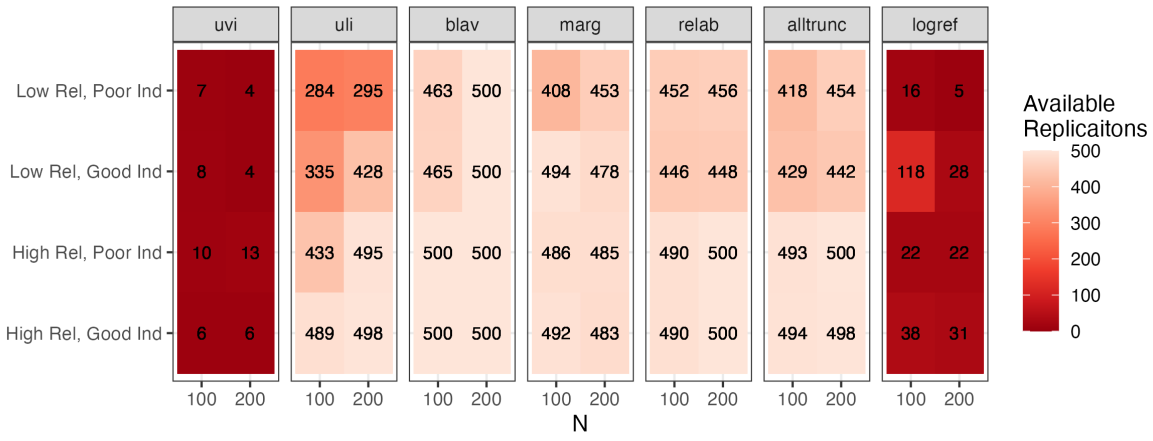
*The percentage of converged runs out of 500 replications.*




*Note.* The criteria are: $\hat{r} < 1.1$, ESS (bulk or tail) $> 300$ (100 per chain), MCSE/MCSD $<$ 10%. MCSD: Monte Carlo standard deviation.

**Figure 2**

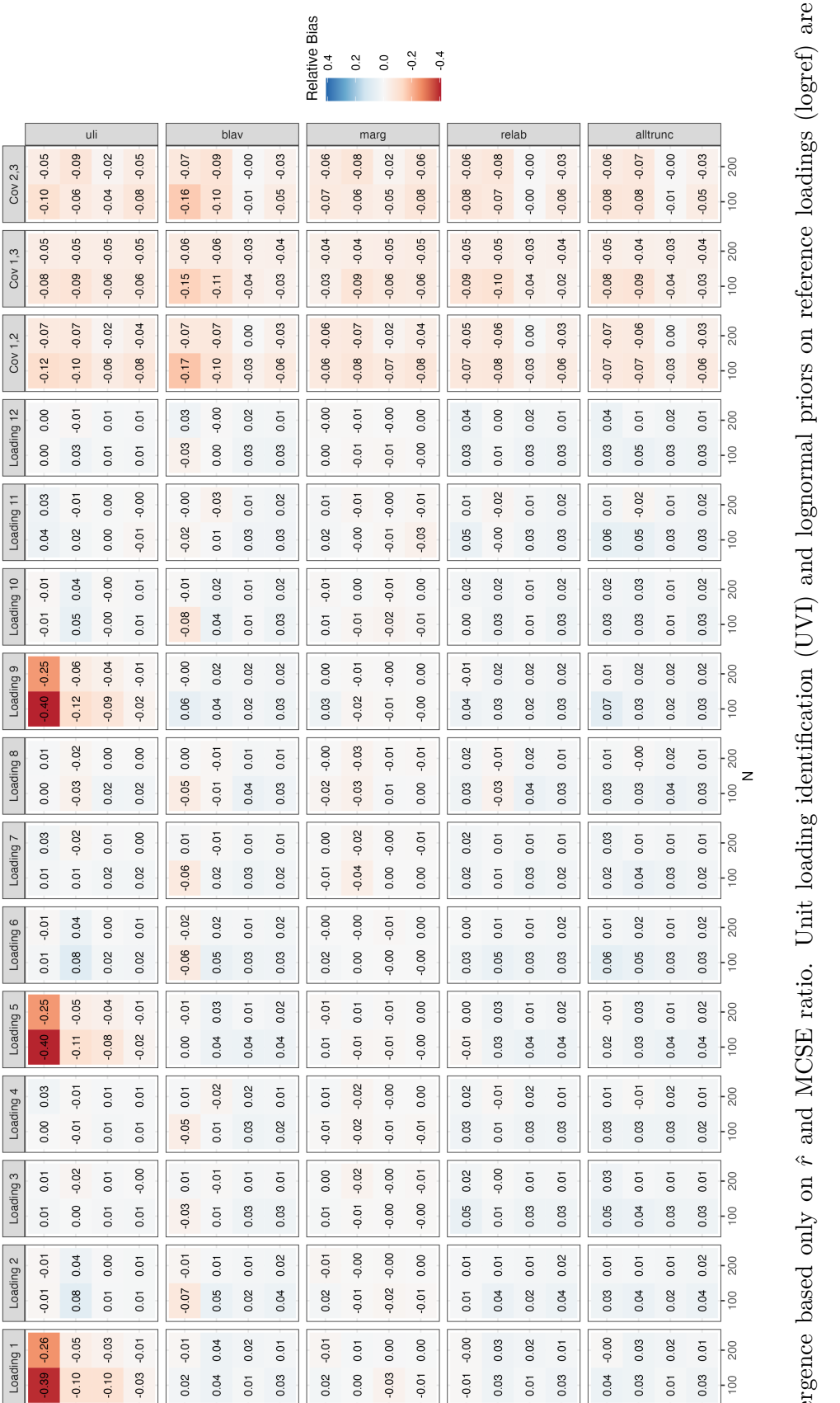
*The number of converged replications.*




*Note.* Convergence based only on $\hat{r}$ and MCSE ratio.

**Figure 3**

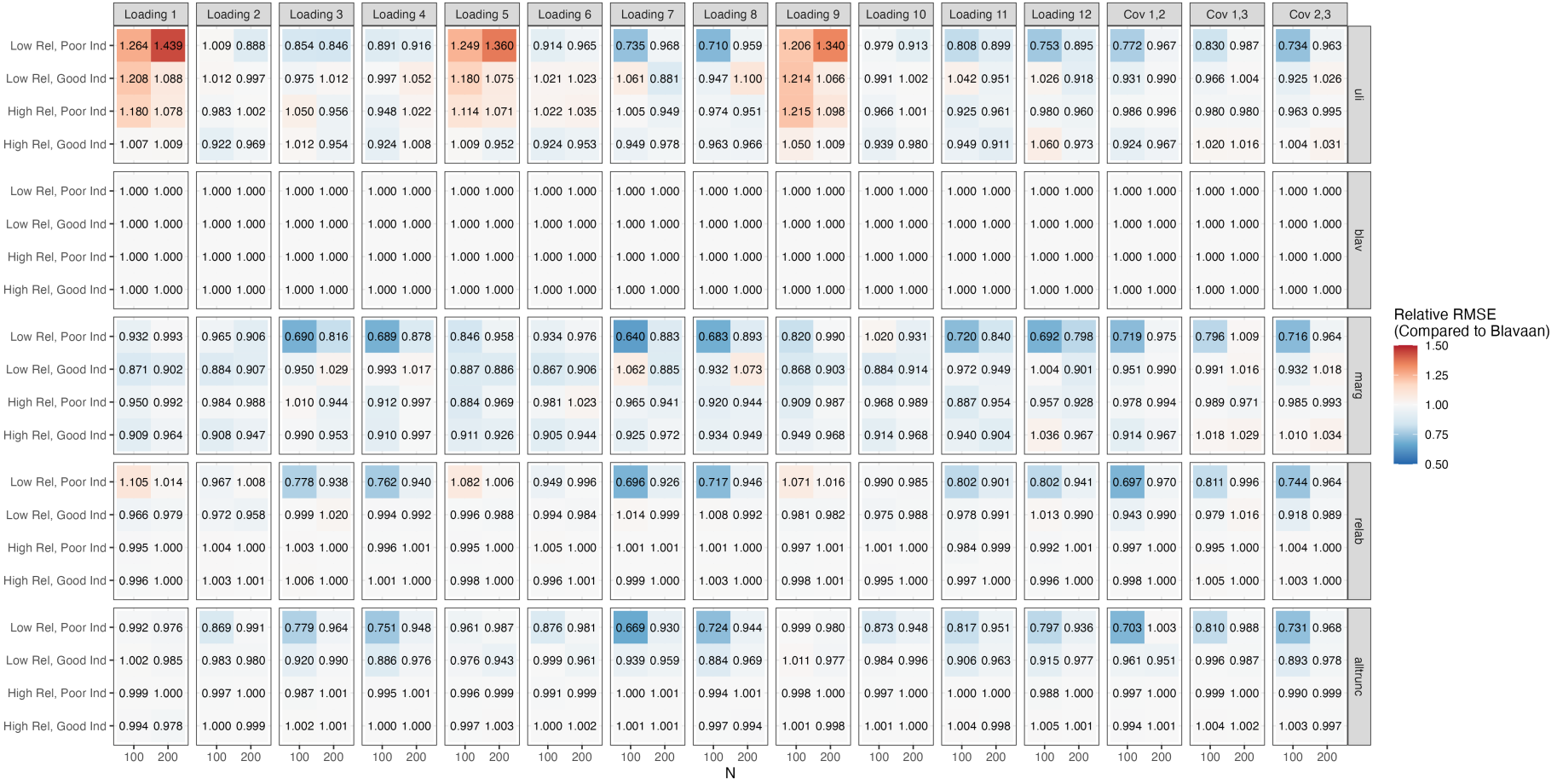
*The average relative bias over the converged replications.*




*Note.* Convergence based only on $\hat{r}$ and MCSE ratio. Unit loading identification (UVI) and lognormal priors on reference loadings (logref) are excluded due to low convergence.

**Figure 4**

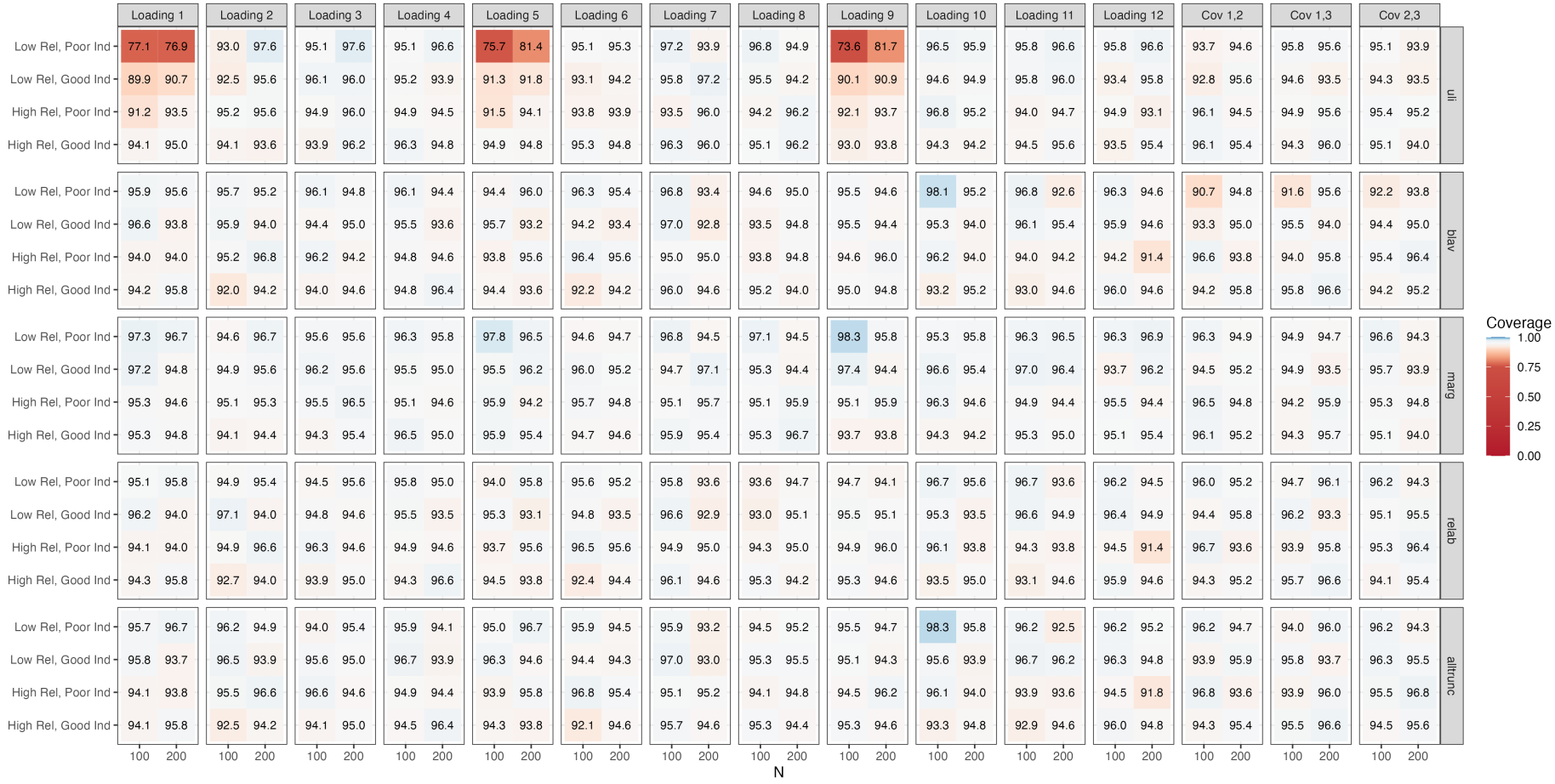
*The relative RMSE over the converged replications compared to the in-iterations relabeling approach in* blavaan.




*Note.* Convergence based only on $\hat{r}$ and MCSE ratio. Unit loading identification (UVI) and lognormal priors on reference loadings (logref) are excluded due to low convergence. Blue denotes lower (better) RMSE than blavaan, where red denotes higher (worse) RMSE than blavaan.

**Figure 5**

*The coverage of the 95% credible interval over the converged replications.*




*Note.* Convergence based only on $\hat{r}$ and MCSE ratio. Unit loading identification (UVI) and lognormal priors on reference loadings (logref) are excluded due to low convergence.

# References

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using mplus: Technical implementation.*

Bollen, K. A., Lilly, A. G., & Luo, L. (2022). Selecting scaling indicators in structural equation models (SEMs). *Psychological methods.*

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors for detecting differential item functioning relative to other bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(1), 122–139.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of econometrics*, *183*(1), 31–57.

Erosheva, E. A., & Curtis, S. M. (2017). Dealing with reflection invariance in bayesian factor analysis. *Psychometrika*, *82*, 295–307.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Hayashi, K., & Marcoulides, G. A. (2006). Teacher's corner: Examining identification issues in factor analysis. *Structural Equation Modeling*, *13*(4), 631–645.

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, *1970*(2), i–41.

Kim, C., Daniels, M., Li, Y., Milbury, K., & Cohen, L. (2018). A bayesian semiparametric latent variable approach to causal mediation. *Statistics in medicine*, *37*(7), 1149–1161.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, *100*(9), 1989–2001.

Liu, H., Heo, I., Depaoli, S., & Ivanov, A. (2025). Parameter recovery for misspecified latent mediation models in the bayesian framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(4), 618–637.

Liu, H., Heo, I., Ivanov, A., & Depaoli, S. (2025). Model assumption violations in

bayesian latent mediation analysis: An exploration of bayesian sem fit indices and ppp. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–31.

Martin-Fernandez, M., & Revuelta, J. (2017). Bayesian estimation of multidimensional item response models. a comparison of analytic and simulation algorithms. *Psicologica: International Journal of Methodology and Experimental Psychology*, *38*(1), 25–55.

Merkle, E. C., Ariyo, O., Winter, S. D., & Garnier-Villarreal, M. (2023). Opaque prior distributions in bayesian latent variable models. *preprint arXiv:2301.08667*.

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient bayesian structural equation modeling in stan. *Journal of Statistical Software*, *100*, 1–22.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. doi: 10.18637/jss.v085.i04

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, *8*(1), 1–17.

Muthén, B., & Asparouhov, T. (2012). *New developments in mplus version 7: Part 1.* Presentation at Utrecht University. Retrieved from https://www.statmodel.

Muthén, L. K., & Muthén, B. O. (2017). Mplus users guide. eighth edition. [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Smid, S. C., & Winter, S. D. (2020). Dangers of the defaults: A tutorial on the impact of default priors when using bayesian sem with small samples. *Frontiers in Psychology*, *11*, 611963.

Stan Development Team. (2025). Stan modeling language users guide and reference manual, version 2.36 [Computer software manual]. Retrieved from https://mc-stan.org

Sun, R., Zhou, X., & Song, X. (2021). Bayesian causal mediation analysis with latent mediators and survival outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 778–790.

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modelinga comparison of constrained maximum likelihood, bayesian estimation, and fixed reliability approaches. *Psychological Methods*, *28*(3), 527.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). Retrieved from http://dx.doi.org/10.1214/20-BA1221 doi: 10.1214/20-ba1221